    b.   Activate biological molecule
    c.   Selective phosphate bond formation without protecting groups
5   Mono/Dioxygenase
    a.   Direct oxyfunctionalization of unactivated organic substrates
    b.   Hydroxylation of alkane, aromatics, steroids
    c.   Epoxidation of alkenes
    d.   Enantioselective sulphoxidation
    e.   Regio- and stereoselective Bayer-Villiger oxidations
6   Haloperoxidase
    a.   Oxidative addition of halide ion to nucleophilic sites
    b.   Addition of hypohalous acids to olefinic bonds
    c.   Ring cleavage of cyclopropanes
    d.   Activated aromatic substrates converted to *ortho* and *para* derivatives
    e.   1.3 diketones converted to 2-halo-derivatives
    f.   Heteroatom oxidation of sulfur and nitrogen containing substrates
    g.   Oxidation of enol acetates, alkynes and activated aromatic rings
7   Lignin peroxidase/Diarylpropane peroxidase
    a.   Oxidative cleavage of C-C bonds
    b.   Oxidation of benzylic alcohols to aldehydes
    c.   Hydroxylation of benzylic carbons
    d.   Phenol dimerization
    e.   Hydroxylation of double bonds to form diols
    f.   Cleavage of lignin aldehydes
8   Epoxide hydrolase
    a.   Synthesis of enantiomerically pure bioactive compounds
    b.   Regio- and enantioselective hydrolysis of epoxide
    c.   Aromatic and olefinic epoxidation by monooxygenases to form epoxides
    d.   Resolution of racemic epoxides
    e.   Hydrolysis of steroid epoxides
9   Nitrile hydratase/nitrilase
    a.   Hydrolysis of aliphatic nitriles to carboxamides
    b.   Hydrolysis of aromatic, heterocyclic, unsaturated aliphatic nitriles to corresponding acids
    c.   Hydrolysis of acrylonitrile
    d.   Production of aromatic and carboxamides, carboxylic acids (nicotinamide, picolinamide, isonicotinamide)
    e.   Regioselective hydrolysis of acrylic dinitrile
    f.   α-amino acids from α-hydroxynitriles
10 Transaminase
    a.   Transfer of amino groups into oxo-acids
11 Amidase/Acylase
    a.   Hydrolysis of amides, amidines, and other C-N bonds
    b.   Non-natural amino acid resolution and synthesis

These exemplifications, while illustrating certain specific aspects of the invention, do not portray the limitations or circumscribe the scope of the disclosed invention.

Thus according to one aspect of this invention, the sequences of a plurality of progenitor nucleic acid templates are aligned in order to select one or more demarcation

5   points, which demarcation points can be located at an area of homology, and are comprised of one or more nucleotides, and which demarcation points are shared by at least two of the progenitor templates. The demarcation points can be used to delineate the boundaries of nucleic acid building blocks to be generated. Thus, the demarcation points identified and selected in the progenitor molecules serve as potential chimerization points in the assembly

10   of the progeny molecules.

A serviceable demarcation point can be an area of homology (comprised of at least one homologous nucleotide base) shared by at least two progenitor templates. More preferably a serviceable demarcation point is an area of homology that is shared by at least half of the progenitor templates. Or a serviceable demarcation point is an area of homology

15   that is shared by at least two thirds of the progenitor templates. Even more preferably a serviceable demarcation points is an area of homology that is shared by at least three fourths of the progenitor templates. Even or a serviceable demarcation points is an area of homology that is shared by almost all of the progenitor templates. Even or a serviceable demarcation point is an area of homology that is shared by all of the progenitor templates.

20   The process of designing nucleic acid building blocks and of designing the mutually compatible ligatable ends of the nucleic acid building blocks to be assembled is illustrated in Figures 6 and 7. As shown, the alignment of a set of progenitor templates reveals several naturally occurring demarcation points, and the identification of demarcation points shared by these templates helps to non-stochastically determine the building blocks to be generated

25   and used for the generation of the progeny chimeric molecules.

In one aspect, this invention provides that the ligation reassembly process is performed exhaustively in order to generate an exhaustive library. In other words, all possible ordered combinations of the nucleic acid building blocks are represented in the set of finalized chimeric nucleic acid molecules. At the same time, in a particularly preferred

30   embodiment, the assembly order (i.e. the order of assembly of each building block in the 5' to 3 sequence of each finalized chimeric nucleic acid) in each combination is by design (or

452

non-stochastic). Because of the non-stochastic nature of this invention, the possibility of unwanted side products is greatly reduced.

In one aspect, this invention provides that, the ligation reassembly process is performed systematically, for example in order to generate a systematically compartmentalized library, with compartments that can be screened systematically, e.g. one by one. In other words this invention provides that, through the selective and judicious use of specific nucleic acid building blocks, coupled with the selective and judicious use of sequentially stepped assembly reactions, an experimental design can be achieved where specific sets of progeny products are made in each of several reaction vessels. This allows a systematic examination and screening procedure to be performed. Thus, it allows a potentially very large number of progeny molecules to be examined systematically in smaller groups.

Because of its ability to perform chimerizations in a manner that is highly flexible yet exhaustive and systematic as well, particularly when there is a low level of homology among the progenitor molecules, the instant invention provides for the generation of a library (or set) comprised of a large number of progeny molecules. Because of the non-stochastic nature of the instant ligation reassembly invention, the progeny molecules generated can comprise a library of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design. In one aspect, such a generated library is comprised of greater than 10^3 different progeny molecular species, or greater than 10^5 different progeny molecular species, or greater than 10^{10} different progeny molecular species, or greater than 10^{15} different progeny molecular species, or greater than 10^{20} different progeny molecular species, or greater than 10^{30} different progeny molecular species, or greater than 10^{40} different progeny molecular species, or greater than 10^{50} different progeny molecular species, or greater than 10^{60} different progeny molecular species, or greater than 10^{70} different progeny molecular species, or greater than 10^{80} different progeny molecular species, or greater than 10^{100} different progeny molecular species, or greater than 10^{110} different progeny molecular species, or greater than 10^{120} different progeny molecular species, or greater than 10^{130} different progeny molecular species, or greater than 10^{140} different progeny molecular species, or greater than 10^{150} different progeny molecular species, or greater than 10^{175} different progeny molecular species, or greater than 10^{200} different progeny molecular species, or greater than 10^{300} different progeny molecular species, or greater than 10^{400}

different progeny molecular species, or greater than 10^{500} different progeny molecular species, or greater than 10^{1000} different progeny molecular species.

In one aspect, a set of finalized chimeric nucleic acid molecules, produced as described is comprised of a polynucleotide encoding a polypeptide. In one aspect, this polynucleotide is a gene, which may be a man-made gene. According to another preferred embodiment, this polynucleotide is a gene pathway, which may be a man-made gene pathway. This invention provides that one or more man-made genes generated by this invention may be incorporated into a man-made gene pathway, such as pathway operable in a eukaryotic organism (including a plant).

It is appreciated that the power of this invention is exceptional, as there is much freedom of choice and control regarding the selection of demarcation points, the size and number of the nucleic acid building blocks, and the size and design of the couplings. It is appreciated, furthermore, that the requirement for intermolecular homology is highly relaxed for the operability of this invention. In fact, demarcation points can even be chosen in areas of little or no intermolecular homology. For example, because of codon wobble, i.e. the degeneracy of codons, nucleotide substitutions can be introduced into nucleic acid building blocks without altering the amino acid originally encoded in the corresponding progenitor template. Alternatively, a codon can be altered such that the coding for an originally amino acid is altered. This inventiop provides that such substitutions can be introduced into the nucleic acid building block in order to increase the incidence of intermolecularly homologous demarcation points and thus to allow an increased number of couplings to be achieved among the building blocks, which in turn allows a greater number of progeny chimeric molecules to be generated.

In another exemplifaction, the synthetic nature of the step in which the building blocks are generated allows the design and introduction of nucleotides (e.g. one or more nucleotides, which may be, for example, codons or introns or regulatory sequences) that can later be optionally removed in an in vitro process (e.g. by mutageneis) or in an in vivo process (e.g. by utilizing the gene splicing ability of a host organism). It is appreciated that in many instances the introduction of these nucleotides may also be desirable for many other reasons in addition to the potential benefit of creating a serviceable demarcation point.

According to another embodiment, this invention provides that a nucleic acid building block can be used to introduce an intron. Thus, this invention provides that functional

454

introns may be introduced into a man-made gene of this invention. This invention also provides that functional introns may be introduced into a man-made gene pathway of this invention. Accordingly, this invention provides for the generation of a chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced intron(s).

Accordingly, this invention also provides for the generation of a chimeric polynucleotide that is a man-made gene pathway containing one (or more) artificially introduced intron(s). Preferably, the artificially introduced intron(s) are functional in one or more host cells for gene splicing much in the way that naturally-occurring introns serve functionally in gene splicing. This invention provides a process of producing man-made intron-containing polynucleotides to be introduced into host organisms for recombination and/or splicing.

The ability to achieve chimerizations, using couplings as described herein, in areas of little or no homology among the progenitor molecules, is particularly useful, and in fact critical, for the assembly of novel gene pathways. This invention thus provides for the generation of novel man-made gene pathways using synthetic ligation reassembly. In a particular aspect, this is achieved by the introduction of regulatory sequences, such as promoters, that are operable in an intended host, to confer operability to a novel gene pathway when it is introduced into the intended host. In a particular exemplification, this invention provides for the generation of novel man-made gene pathways that is operable in a plurality of intended hosts (e.g. in a microbial organism as well as in a plant cell). This can be achieve, for example, by the introduction of a plurality of regulatory sequences, comprised of a regulatory sequence that is operable in a first intended host and a regulatory sequence that is operable in a second intended host. A similar process can be performed to achieve operability of a gene pathway in a third intended host species, etc. The number of intended host species can be each integer from 1 to 10 or alternatively over 10. Alternatively, for example, operability of a gene pathway in a plurality of intended hosts can be achieved by the introduction of a regulatory sequence having intrinsic operability in a plurality of intended hosts.

In one aspect, this invention provides that a nucleic acid building block can be used to introduce a regulatory sequence, particularly a regulatory sequence for gene expression. Preferred regulatory sequences include, but are not limited to, those that are man-made, and

those found in archeal, bacterial, eukaryotic (including mitochondrial), viral, and prionic or prion-like organisms. Preferred regulatory sequences include but are not limited to, promoters, operators, and activator binding sites. Thus, this invention provides that functional regulatory sequences may be introduced into a man-made gene of this invention. This invention also provides that functional regulatory sequences may be introduced into a man-made gene pathway of this invention.

Accordingly, this invention provides for the generation of a chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced regulatory sequence(s). Accordingly, this invention also provides for the generation of a chimeric polynucleotide that is a man-made gene pathway containing one (or more) artificially introduced regulatory sequence(s). Preferably, an artificially introduced regulatory sequence(s) is operatively linked to one or more genes in the man-made polynucleotide, and are functional in one or more host cells.

Exemplary bacterial promoters that are serviceable for this invention include lacI, lacZ, T3, T7, gpt, lambda P_R, P_L and trp. Serviceable eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Particular plant regulatory sequences include promoters active in directing transcription in plants, either constitutively or stage and/or tissue specific, depending on the use of the plant or parts thereof. These promoters include, but are not limited to promoters showing constitutive expression, such as the 35S promoter of Cauliflower Mosaic Virus (CaMV) (Guilley et al., 1982), those for leaf-specific expression, such as the promoter of the ribulose bisphosphate carboxylase small subunit gene (Coruzzi et al., 1984), those for root-specific expression, such as the promoter from the glutamin synthase gene (Tingey et al., 1987), those for seed-specific expression, such as the cruciferin A promoter from *Brassica napus* (Ryan et al., 1989), those for tuber-specific expression, such as the class-I patatin promoter from potato (Rocha-Sasa et al., 1989; Wenzler et al., 1989) or those for fruit-specific expression, such as the polygalacturonase (PG) promoter from tomato (Bird et al., 1988).

Other regulatory sequences that are preferred for this invention include terminator sequences and polyadenylation signals and any such sequence functioning as such in plants, the choice of which is within the level of the skilled artisan. An example of such sequences

456

is the 3' flanking region of the nopaline synthase (nos) gene of *Agrobacterium tumefaciens* (Bevan, 1984). The regulatory sequences may also include enhancer sequences, such as found in the 35S promoter of CaMV, and mRNA stabilizing sequences such as the leader sequence of Alfalfa Mosaic Cirus (AlMV) RNA4 (Brederode et al., 1980) or any other

5   sequences functioning in a like manner.

A man-made genes produced using this invention can also serve as a substrate for recombination with another nucleic acid. Likewise, a man-made gene pathway produced using this invention can also serve as a substrate for recombination with another nucleic acid. In a preferred instance, the recombination is facilitated by, or occurs at, areas of homology

10   between the man-made intron-containing gene and a nucleic acid with serves as a recombination partner. In a particularly preferred instance, the recombination partner may also be a nucleic acid generated by this invention, including a man-made gene or a man-made gene pathway. Recombination may be facilitated by or may occur at areas of homology that exist at the one (or more) artificially introduced intron(s) in the man-made gene.

15   The synthetic ligation reassembly method of this invention utilizes a plurality of nucleic acid building blocks, each of which preferably has two ligatable ends. The two ligatable ends on each nucleic acid building block may be two blunt ends (i.e. each having an overhang of zero nucleotides), or preferably one blunt end and one overhang, or or two overhangs.

20   A serviceable overhang for this purpose may be a 3' overhang or a 5' overhang. Thus, a nucleic acid building block may have a 3' overhang or alternatively a 5' overhang or alternatively two 3' overhangs or alternatively two 5' overhangs. The overall order in which the nucleic acid building blocks are assembled to form a finalized chimeric nucleic acid molecule is determined by purposeful experimental design and is not random.

25   In one aspect, a nucleic acid building block is generated by chemical synthesis of two single-stranded nucleic acids (also referred to as single-stranded oligos) and contacting them so as to allow them to anneal to form a double-stranded nucleic acid building block.

A double-stranded nucleic acid building block can be of variable size. The sizes of these building blocks can be small or large depending on the choice of the experimenter.

30   Preferred sizes for building block range from 1 base pair (not including any overhangs) to 100,000 base pairs (not including any overhangs). Other preferred size ranges are also

provided, which have lower limits of from 1 bp to 10,000 bp (including every integer value in between), and upper limits of from 2 bp to 100, 000 bp (including every integer value in between).

It is appreciated that current methods of polymerase-based amplification can be used to generate double-stranded nucleic acids of up to thousands of base pairs, if not tens of thousands of base pairs, in length with high fidelity. Chemical synthesis (e.g. phosphoramidite-based) can be used to generate nucleic acids of up to hundreds of nucleotides in length with high fidelity; however, these can be assembled, e.g. using overhangs or sticky ends, to form double-stranded nucleic acids of up to thousands of base pairs, if not tens of thousands of base pairs, in length if so desired.

A combination of methods (e.g. phosphoramidite-based chemical synthesis and PCR) can also be used according to this invention. Thus, nucleic acid building block made by different methods can also be used in combination to generate a progeny molecule of this invention.

The use of chemical synthesis to generate nucleic acid building blocks is particularly preferred in this invention & is advantageous for other reasons as well, including procedural safety and ease. No cloning or harvesting or actual handling of any biological samples is required. The design of the nucleic acid building blocks can be accomplished on paper. Accordingly, this invention teaches an advance in procedural safety in recombinant technologies.

In one aspect, a double-stranded nucleic acid building block according to this invention may also be generated by polymerase-based amplification of a polynucleotide template. In a non-limiting exemplification, as illustrated in Figure 2, a first polymerase-based amplification reaction using a first set of primers, F_2 and R_1, is used to generate a blunt-ended product (labeled Reaction 1, Product 1), which is essentially identical to Product A. A second polymerase-based amplification reaction using a second set of primers, F_1 and R_2, is used to generate a blunt-ended product (labeled Reaction 2, Product 2), which is essentially identical to Product B. These two products are mixed and allowed to melt and anneal, generating potentially useful double-stranded nucleic acid building blocks with two overhangs. In the example of Fig. 2, the product with the 3' overhangs (Product C) is selected by nuclease-based degradation of the other 3 products using a 3' acting exonuclease, such as exonuclease III. It is appreciated that a 5' acting exonuclease (e.g. red alpha) may be

458

also be used, for example to select Product D instead. It is also appreciated that other selection means can also be used, including hybridization-based means, and that these means can incorporate a further means, such as a magnetic bead-based means, to facilitate separation of the desired product.

5   Many other methods exist by which a double-stranded nucleic acid building block can be generated that is serviceable for this invention; and these are known in the art and can be readily performed by the skilled artisan.

In one aspect, a double-stranded nucleic acid building block that is serviceable for this invention is generated by first generating two single stranded nucleic acids and allowing

10   them to anneal to form a double-stranded nucleic acid building block. The two strands of a double-stranded nucleic acid building block may be complementary at every nucleotide apart from any that form an overhang; thus containing no mismatches, apart from any overhang(s). According to another embodiment, the two strands of a double-stranded nucleic acid building block are complementary at fewer than every nucleotide apart from any that form an

15   overhang. Thus, according to this embodiment, a double-stranded nucleic acid building block can be used to introduce codon degeneracy. Preferably the codon degeneracy is introduced using the site-saturation mutagenesis described herein, using one or more N,N,G/T cassettes or alternatively using one or more N,N,N cassettes.

Contained within an exemplary experimental design for achieving an ordered

20   assembly according to this invention are:

1) The design of specific nucleic acid building blocks.

2) The design of specific ligatable ends on each nucleic acid building block.

3) The design of a particular order of assembly of the nucleic acid building blocks.

An overhang may be a 3' overhang or a 5' overhang. An overhang may also have a

25   terminal phosphate group or alternatively may be devoid of a terminal phosphate group (having, e.g., a hydroxyl group instead). An overhang may be comprised of any number of nucleotides. Preferably an overhang is comprised of 0 nucleotides (as in a blunt end) to 10,000 nucleotides. Thus, a wide range of overhang sizes may be serviceable. Accordingly, the lower limit may be each integer from 1-200 and the upper limit may be each integer from

30   2-10,000. According to a particular exemplification, an overhang may consist of anywhere from 1 nucleotide to 200 nucleotides (including every integer value in between).

The final chimeric nucleic acid molecule may be generated by sequentially assembling 2 or more building blocks at a time until all the designated building blocks have been assembled. A working sample may optionally be subjected to a process for size selection or purification or other selection or enrichment process between the performance of

5   two assembly steps. Alternatively, the final chimeric nucleic acid molecule may be generated by assembling all the designated building blocks at once in one step.

Utility

The *in vivo* recombination method of this invention can be performed blindly on a pool of unknown hybrids or alleles of a specific polynucleotide or sequence. However, it is

10   not necessary to know the actual DNA or RNA sequence of the specific polynucleotide.

The approach of using recombination within a mixed population of genes can be useful for the generation of any useful proteins, for example, interleukin I, antibodies, tPA and growth hormone. This approach may be used to generate proteins having altered specificity or activity. The approach may also be useful for the generation of hybrid nucleic

15   acid sequences, for example, promoter regions, introns, exons, enhancer sequences, 31 untranslated regions or 51 untranslated regions of genes. Thus this approach may be used to generate genes having increased rates of expression. This approach may also be useful in the study of repetitive DNA sequences. Finally, this approach may be useful to mutate ribozymes or aptamers.

20   Scaffold-like regions separating regions of diversity in proteins may be particularly suitable for the methods of this invention. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta barrel, and the four-helix bundle. The methods of this invention can be used to create scaffold-like proteins

25   with various combinations of mutated sequences for binding.

The equivalents of some standard genetic matings may also be performed by the methods of this invention. For example, a "molecular" backcross can be performed by repeated mixing of the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to

30   combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (i.e.

immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not.

End-Selection

This invention provides a method for selecting a subset of polynucleotides from a starting set of polynucleotides, which method is based on the ability to discriminate one or more selectable features (or selection markers) present anywhere in a working polynucleotide, so as to allow one to perform selection for (positive selection) &/or against (negative selection) each selectable polynucleotide. In a preferred aspect, a method is provided termed end-selection, which method is based on the use of a selection marker located in part or entirely in a terminal region of a selectable polynucleotide, and such a selection marker may be termed an "end-selection marker".

End-selection may be based on detection of naturally occurring sequences or on detection of sequences introduced experimentally (including by any mutagenesis procedure mentioned herein and not mentioned herein) or on both, even within the same polynucleotide. An end-selection marker can be a structural selection marker or a functional selection marker or both a structural and a functional selection marker. An end-selection marker may be comprised of a polynucleotide sequence or of a polypeptide sequence or of any chemical structure or of any biological or biochemical tag, including markers that can be selected using methods based on the detection of radioactivity, of enzymatic activity, of fluorescence, of any optical feature, of a magnetic property (e.g. using magnetic beads), of immunoreactivity, and of hybridization.

End-selection may be applied in combination with any method serviceable for performing mutagenesis. Such mutagenesis methods include, but are not limited to, methods described herein (supra and infra). Such methods include, by way of non-limiting exemplification, any method that may be referred herein or by others in the art by any of the following terms: "saturation mutagenesis", "shuffling", "recombination", "re-assembly", "error-prone PCR", "assembly PCR", "sexual PCR", "crossover PCR", "oligonucleotide primer-directed mutagenesis", "recursive (&/or exponential) ensemble mutagenesis (see Arkin and Youvan, 1992)", "cassette mutagenesis", "in vivo mutagenesis", and "in vitro mutagenesis". Moreover, end-selection may be performed on molecules produced by any mutagenesis &/or amplification method (see, e.g., Arnold, 1993; Caldwell and Joyce, 1992;

Stemmer, 1994; following which method it is desirable to select for (including to screen for the presence of) desirable progeny molecules.

In addition, end-selection may be applied to a polynucleotide apart from any mutagenesis method. In a preferred embodiment, end-selection, as provided herein, can be used in order to facilitate a cloning step, such as a step of ligation to another polynucleotide (including ligation to a vector). This invention thus provides for end-selection as a serviceable means to facilitate library construction, selection &/or enrichment for desirable polynucleotides, and cloning in general.

In a particularly preferred embodiment, end-selection can be based on (positive) selection for a polynucleotide; alternatively end-selection can be based on (negative) selection against a polynucleotide; and alternatively still, end-selection can be based on both (positive) selection for, and on (negative) selection against, a polynucleotide. End-selection, along with other methods of selection &/or screening, can be performed in an iterative fashion, with any combination of like or unlike selection &/or screening methods and serviceable mutagenesis methods, all of which can be performed in an iterative fashion and in any order, combination, and permutation.

It is also appreciated that, according to one embodiment of this invention, end-selection may also be used to select a polynucleotide is at least in part: circular (e.g. a plasmid or any other circular vector or any other polynucleotide that is partly circular), &/or branched, &/or modified or substituted with any chemical group or moiety. In accord with this embodiment, a polynucleotide may be a circular molecule comprised of an intermediate or central region, which region is flanked on a 5' side by a 5' flanking region (which, for the purpose of end-selection, serves in like manner to a 5' terminal region of a non-circular polynucleotide) and on a 3' side by a 3' terminal region (which, for the purpose of end-selection, serves in like manner to a 3' terminal region of a non-circular polynucleotide). As used in this non-limiting exemplification, there may be sequence overlap between any two regions or even among all three regions.

In one non-limiting aspect of this invention, end-selection of a linear polynucleotide is performed using a general approach based on the presence of at least one end-selection marker located at or near a polynucleotide end or terminus (that can be either a 5' end or a 3' end). In one particular non-limiting exemplification, end-selection is based on selection for a specific sequence at or near a terminus such as, but not limited to, a sequence recognized by

462

an enzyme that recognizes a polynucleotide sequence. An enzyme that recognizes and catalyzes a chemical modification of a polynucleotide is referred to herein as a polynucleotide-acting enzyme. In a preferred embodiment, serviceable polynucleotide-acting enzymes are exemplified non-exclusively by enzymes with polynucleotide-cleaving activity, enzymes with polynucleotide-methylating activity, enzymes with polynucleotide-ligating activity, and enzymes with a plurality of distinguishable enzymatic activities (including non-exclusively, e.g., both polynucleotide-cleaving activity and polynucleotide-ligating activity).

Relevant polynucleotide-acting enzymes thus also include any commercially available or non-commercially available polynucleotide endonucleases and their companion methylases, see, e.g., Roberts and Macelis, 1996). Preferred polynucleotide endonucleases include – but are not limited to – type II restriction enzymes (including type IIS), and include enzymes that cleave both strands of a double stranded polynucleotide (e.g. Not I, which cleaves both strands at 5'...GC/GGCCGC...3') and enzymes that cleave only one strand of a double stranded polynucleotide, i.e. enzymes that have polynucleotide-nicking activity, (e.g. N. BstNB I, which cleaves only one strand at 5'...GAGTCNNNN/N...3'). Relevant polynucleotide-acting enzymes also include type III restriction enzymes.

It is appreciated that relevant polynucleotide-acting enzymes also include any enzymes that may be developed in the future, though currently unavailable, that are serviceable for generating a ligation compatible end, preferably a sticky end, in a polynucleotide.

In one exemplification, a serviceable selection marker is a restriction site in a polynucleotide that allows a corresponding type II (or type IIS) restriction enzyme to cleave an end of the polynucleotide so as to provide a ligatable end (including a blunt end or alternatively a sticky end with at least a one base overhang) that is serviceable for a desirable ligation reaction without cleaving the polynucleotide internally in a manner that destroys a desired internal sequence in the polynucleotide. Thus it is provided that, among relevant restriction sites, those sites that do not occur internally (i.e. that do not occur apart from the termini) in a specific working polynucleotide are preferred when the use of a corresponding restriction enzyme(s) is not intended to cut the working polynucleotide internally. This allows one to perform restriction digestion reactions to completion or to near completion without incurring unwanted internal cleavage in a working polynucleotide.

According to a preferred aspect, it is thus preferable to use restriction sites that are not contained, or alternatively that are not expected to be contained, or alternatively that unlikely to be contained (e.g. when sequence information regarding a working polynucleotide is incomplete) internally in a polynucleotide to be subjected to end-selection. In accordance

5   with this aspect, it is appreciated that restriction sites that occur relatively infrequently are usually preferred over those that occur more frequently. On the other hand it is also appreciated that there are occasions where internal cleavage of a polypeptide is desired, e.g. to achieve recombination or other mutagenic procedures along with end-selection.

In accord with this invention, it is also appreciated that methods (e.g. mutagenesis

10   methods) can be used to remove unwanted internal restriction sites. It is also appreciated that a partial digestion reaction (i.e. a digestion reaction that proceeds to partial completion) can be used to achieve digestion at a recognition site in a terminal region while sparing a susceptible restriction site that occurs internally in a polynucleotide and that is recognized by the same enzyme. In one aspect, partial digest are useful because it is appreciated that certain

15   enzymes show preferential cleavage of the same recognition sequence depending on the location and environment in which the recognition sequence occurs. For example, it is appreciated that, while lambda DNA has 5 *EcoR* I sites, cleavage of the site nearest to the right terminus has been reported to occur 10 times faster than the sites in the middle of the molecule. Also, for example, it has been reported that, while *Sac* II has four sites on lambda

20   DNA, the three clustered centrally in lambda are cleaved 50 times faster than the remaining site near the terminus (at nucleotide 40,386). Summarily, site preferences have been reported for various enzymes by many investigators (e.g., Thomas and Davis, 1975; Forsblum et al, 1976; Nath and Azzolina, 1981; Brown and Smith, 1977; Gingeras and Brooks, 1983; Krüger et al, 1988; Conrad and Topal, 1989; Oller et al, 1991; Topal, 1991; and Pein, 1991; to name

25   but a few). It is appreciated that any empirical observations as well as any mechanistic understandings of site preferences by any serviceable polynucleotide-acting enzymes, whether currently available or to be procured in the future, may be serviceable in end-selection according to this invention.

It is also appreciated that protection methods can be used to selectively protect

30   specified restriction sites (e.g. internal sites) against unwanted digestion by enzymes that would otherwise cut a working polypeptide in response to the presence of those sites; and that such protection methods include modifications such as methylations and base

substitutions (e.g. U instead of T) that inhibit an unwanted enzyme activity. It is appreciated that there are limited numbers of available restriction enzymes that are rare enough (e.g. having very long recognition sequences) to create large (e.g. megabase-long) restriction fragments, and that protection approaches (e.g. by methylation) are serviceable for increasing the rarity of enzyme cleavage sites. The use of M.*Fnu* II (mCGCG) to increase the apparent rarity of *Not* I approximately twofold is but one example among many (Qiang et al, 1990; Nelson et al, 1984; Maxam and Gilbert, 1980; Raleigh and Wilson, 1986).

According to a preferred aspect of this invention, it is provided that, in general, the use of rare restriction sites is preferred. It is appreciated that, in general, the frequency of occurrence of a restriction site is determined by the number of nucleotides contained therein, as well as by the ambiguity of the base requirements contained therein. Thus, in a non-limiting exemplification, it is appreciated that, in general, a restriction site composed of, for example, 8 specific nucleotides (e.g. the *Not* I site or GC/GGCCGC, with an estimated relative occurrence of 1 in 4^8, i.e. 1 in 65,536, random 8-mers) is relatively more infrequent than one composed of, for example, 6 nucleotides (e.g. the *Sma* I site or CCC/GGG, having an estimated relative occurrence of 1 in 4^6, i.e. 1 in 4,096, random 6-mers), which in turn is relatively more infrequent than one composed of, for example, 4 nucleotides (e.g. the *Msp* I site or C/CGG, having an estimated relative occurrence of 1 in 4^4, i.e. 1 in 256, random 4-mers). Moreover, in another non-limiting exemplification, it is appreciated that, in general, a restriction site having no ambiguous (but only specific) base requirements (e.g. the *Fin* I site or GTCCC, having an estimated relative occurrence of 1 in 4^5, i.e. 1 in 1024, random 5-mers) is relatively more infrequent than one having an ambiguous W (where W = A or T) base requirement (e.g. the *Ava* II site or G/GWCC, having an estimated relative occurrence of 1 in 4x4x2x4x4 – i.e. 1 in 512 – random 5-mers), which in turn is relatively more infrequent than one having an ambiguous N (where N = A or C or G or T) base requirement (e.g. the *Asu* I site or G/GNCC, having an estimated relative occurrence of 1 in 4x4x1x4x4, i.e. 1 in 256 – random 5-mers). These relative occurrences are considered general estimates for actual polynucleotides, because it is appreciated that specific nucleotide bases (not to mention specific nucleotide sequences) occur with dissimilar frequencies in specific polynucleotides, in specific species of organisms, and in specific groupings of organisms. For example, it is appreciated that the % G+C contents of different species of organisms are often very different and wide ranging.

The use of relatively more infrequent restriction sites as a selection marker include - in a non-limiting fashion - preferably those sites composed at least a 4 nucleotide sequence, more preferably those composed at least a 5 nucleotide sequence, or those composed at least a 6 nucleotide sequence (e.g. the BamH I site or G/GATCC, the Bgl II site or A/GATCT, the Pst I site or CTGCA/G, and the Xba I site or T/CTAGA), or those composed at least a 7 nucleotide sequence, or those composed of an 8 nucleotide sequence nucleotide sequence (e.g. the Asc I site or GG/CGCGCC, the Not I site or GC/GGCCGC, the Pac I site or TTAAT/TAA, the Pme I site or GTTT/AAAC, the Srf I site or GCCC/GGGC, the Sse838 I site or CCTGCA/GG, and the Swa I site or ATTT/AAAT), or those composed of a 9 nucleotide sequence, and even or those composed of at least a 10 nucleotide sequence (e.g. the BspG I site or CG/CGCTGGAC). It is further appreciated that some restriction sites (e.g. for class IIS enzymes) are comprised of a portion of relatively high specificity (i.e. a portion containing a principal determinant of the frequency of occurrence of the restriction site) and a portion of relatively low specificity; and that a site of cleavage may or may not be contained within a portion of relatively low specificity. For example, in the Eco57 I site or CTGAAG(16/14), there is a portion of relatively high specificity (i.e. the CTGAAG portion) and a portion of relatively low specificity (i.e. the N16 sequence) that contains a site of cleavage.

In another embodiment of this invention, a serviceable end-selection marker is a terminal sequence that is recognized by a polynucleotide-acting enzyme that recognizes a specific polynucleotide sequence. In one aspect of this invention, serviceable polynucleotide-acting enzymes also include other enzymes in addition to classic type II restriction enzymes. According to this aspect of this invention, serviceable polynucleotide-acting enzymes also include gyrases, helicases, recombinases, relaxases, and any enzymes related thereto.

Among examples are topoisomerases (which have been categorized by some as a subset of the gyrases) and any other enzymes that have polynucleotide-cleaving activity (including preferably polynucleotide-nicking activity) &/or polynucleotide-ligating activity. Among preferred topoisomerase enzymes are topoisomerase I enzymes, which is available from many commercial sources (Epicentre Technologies, Madison, WI; Invitrogen, Carlsbad, CA; Life Technologies, Gathesburg, MD) and conceivably even more private sources. It is appreciated that similar enzymes may be developed in the future that are serviceable for end-

selection as provided herein. A particularly preferred topoisomerase I enzyme is a topoisomerase I enzyme of vaccinia virus origin, that has a specific recognition sequence (e.g. 5'...AAGGG...3') and has both polynucleotide-nicking activity and polynucleotide-ligating activity. Due to the specific nicking-activity of this enzyme (cleavage of one strand),

5   internal recognition sites are not prone to polynucleotide destruction resulting from the nicking activity (but rather remain annealed) at a temperature that causes denaturation of a terminal site that has been nicked. Thus for use in end-selection, it is preferable that a nicking site for topoisomerase-based end-selection be no more than 100 nucleotides from a terminus, more preferably no more than 50 nucleotides from a terminus, or no more than 25

10   nucloetides from a terminus, even or no more than 20 nucleotides from a terminus, even or no more than 15 nucleotides from a terminus, even or no more than 10 nucleotides from a terminus, even or no more than 8 nucleotides from a terminus, even or no more than 6 nucleotides from a terminus, and even or no more than 4 nucleotides from a terminus.

       In a particularly preferred exemplification that is non-limiting yet clearly illustrative,

15   it is appreciated that when a nicking site for topoisomerase-based end-selection is 4 nucleotides from a terminus, nicking produces a single stranded oligo of 4 bases (in a terminal region) that can be denatured from its complementary strand in an end-selectable polynucleotide; this provides a sticky end (comprised of 4 bases) in a polynucleotide that is serviceable for an ensuing ligation reaction. To accomplish ligation to a cloning vector

20   (preferably an expression vector), compatible sticky ends can be generated in a cloning vector by any means including by restriction enzyme-based means. The terminal nucleotides (comprised of 4 terminal bases in this specific example) in an end-selectable polynucleotide terminus are thus wisely chosen to provide compatibility with a sticky end generated in a cloning vector to which the polynucleotide is to be ligated.

25   On the other hand, internal nicking of an end-selectable polynucleotide, e.g. 500 bases from a terminus, produces a single stranded oligo of 500 bases that is not easily denatured from its complementary strand, but rather is serviceable for repair (e.g. by the same topoisomerase enzyme that produced the nick).

       This invention thus provides a method - e.g. that is vaccinia topoisomerase-based

30   &/or type II (or IIS) restriction endonuclease-based &/or type III restriction endonuclease-based &/or nicking enzyme-based (e.g. using N. BstNB I) – for producing a sticky end in a working polynucleotide, which end is ligation compatible, and which end can be comprised

of at least a 1 base overhang. Preferably such a sticky end is comprised of at least a 2-base overhang, more preferably such a sticky end is comprised of at least a 3-base overhang, or such a sticky end is comprised of at least a 4-base overhang, even or such a sticky end is comprised of at least a 5-base overhang, even or such a sticky end is comprised of at least a

5   6-base overhang. Such a sticky end may also be comprised of at least a 7-base overhang, or at least an 8-base overhang, or at least a 9-base overhang, or at least a 10-base overhang, or at least 15-base overhang, or at least a 20-base overhang, or at least a 25-base overhang, or at least a 30-base overhang. These overhangs can be comprised of any bases, including A, C, G, or T.

10   It is appreciated that sticky end overhangs introduced using topoisomerase or a nicking enzyme (e.g. using N. *Bst*NB I) can be designed to be unique in a ligation environment, so as to prevent unwanted fragment reassemblies, such as self-dimerizations and other unwanted concatamerizations.

           According to one aspect of this invention, a plurality of sequences (which may but do

15   not necessarily overlap) can be introduced into a terminal region of an end-selectable polynucleotide by the use of an oligo in a polymerase-based reaction. In a relevant, but by no means limiting example, such an oligo can be used to provide a preferred 5' terminal region that is serviceable for topoisomerase I-based end-selection, which oligo is comprised of: a 1-10 base sequence that is convertible into a sticky end (preferably by a vaccinia

20   topoisomerase I), a ribosome binding site (i.e. and "RBS", that is preferably serviceable for expression cloning), and optional linker sequence followed by an ATG start site and a template-specific sequence of 0-100 bases (to facilitate annealment to the template in the a polymerase-based reaction). Thus, according to this example, a serviceable oligo (which may be termed a forward primer) can have the sequence: 5'[terminal sequence = $(N)_{1-}$

25    $_{10}$][topoisomerase I site & RBS = AAGGGAGGAG][linker = $(N)_{1-100}$][start codon and template-specific sequence = ATG$(N)_{0-100}$]3'.

           Analogously, in a relevant, but by no means limiting example, an oligo can be used to provide a preferred 3' terminal region that is serviceable for topoisomerase I-based end-selection, which oligo is comprised of: a 1-10 base sequence that is convertible into a sticky

30   end (preferably by a vaccinia topoisomerase I), and optional linker sequence followed by a template-specific sequence of 0-100 bases (to facilitate annealment to the template in the a polymerase-based reaction). Thus, according to this example, a serviceable oligo (which

may be termed a reverse primer) can have the sequence: 5'[terminal sequence = $(N)_{1-10}$][topoisomerase I site = AAGGG][linker = $(N)_{1-100}$][template-specific sequence = $(N)_{0-100}$]3'.

It is appreciated that, end-selection can be used to distinguish and separate parental template molecules (e.g. to be subjected to mutagenesis) from progeny molecules (e.g. generated by mutagenesis). For example, a first set of primers, lacking in a topoisomerase I recognition site, can be used to modify the terminal regions of the parental molecules (e.g. in polymerase-based amplification). A different second set of primers (e.g. having a topoisomerase I recognition site) can then be used to generate mutated progeny molecules (e.g. using any polynucleotide chimerization method, such as interrupted synthesis, template-switching polymerase-based amplification, or interrupted synthesis; or using saturation mutagenesis; or using any other method for introducing a topoisomerase I recognition site into a mutagenized progeny molecule as disclosed herein) from the amplified template molecules. The use of topoisomerase I-based end-selection can then facilitate, not only discernment, but selective topoisomerase I-based ligation of the desired progeny molecules.

Annealment of a second set of primers to thusly amplified parental molecules can be facilitated by including sequences in a first set of primers (i.e. primers used for amplifying a set parental molecules) that are similar to a toposiomerase I recognition site, yet different enough to prevent functional toposiomerase I enzyme recognition. For example, sequences that diverge from the AAGGG site by anywhere from 1 base to all 5 bases can be incorporated into a first set of primers (to be used for amplifying the parental templates prior to subjection to mutagenesis). In a specific, but non-limiting aspect, it is thus provided that a parental molecule can be amplified using the following exemplary – but by no means limiting – set of forward and reverse primers:

Forward Primer: 5' CTAGAAGAGAGGAGAAAACCATG$(N)_{10-100}$ 3', and

Reverse Primer: 5' GATCAAAGGCGCGCCTGCAGG$(N)_{10-100}$ 3'

According to this specific example of a first set of primers, $(N)_{10-100}$ represents preferably a 10 to 100 nucleotide-long template-specific sequence, more preferably a 10 to 50 nucleotide-long template-specific sequence, or a 10 to 30 nucleotide-long template-specific sequence, and even or a 15 to 25 nucleotide-long template-specific sequence.

According to a specific, but non-limiting aspect, it is thus provided that, after this amplification (using a disclosed first set of primers lacking in a true topoisomerase I recognition site), amplified parental molecules can then be subjected to mutagenesis using one or more sets of

forward and reverse primers that do have a true topoisomerase I recognition site. In a specific, but non-limiting aspect, it is thus provided that a parental molecule can be used as templates for the generation of a mutagenized progeny molecule using the following exemplary – but by no means limiting – second set of forward and reverse primers:

5   Forward Primer: 5' CTAGAAGGGAGGAGAAAACCATG 3'

           Reverse Primer: 5' GATCAAAGGCGCGCCTGCAGG 3'   (contains *Asc* I recognition sequence)

           It is appreciated that any number of different primers sets not specifically mentioned can be used as first, second, or subsequent sets of primers for end-selection consistent with this

0   invention. Notice that type II restriction enzyme sites can be incorporated (e.g. an *Asc* I site in the above example).   It is provided that, in addition to the other sequences mentioned, the experimentalist can incorporate one or more N,N,G/T triplets into a serviceable primer in order to subject a working polynucleotide to saturation mutagenesis. Summarily, use of a second and/or subsequent set of primers can achieve dual goals of introducing a topoisomerase I site and of

5   generating mutations in a progeny polynucleotide.

           Thus, according to one use provided, a serviceable end-selection marker is an enzyme recognition site that allows an enzyme to cleave (including nick) a polynucleotide at a specified site, to produce a ligation-compatible end upon denaturation of a generated single stranded oligo. Ligation of the produced polynucleotide end can then be accomplished by

!0   the same enzyme (e.g. in the case of vaccinia virus topoisomerase I), or alternatively with the use of a different enzyme. According to one aspect of this invention, any serviceable end-selection markers, whether like (e.g. two vaccinia virus topoisomerase I recognition sites) or unlike (e.g. a class II restriction enzyme recognition site and a vaccinia virus topoisomerase I recognition site) can be used in combination to select a polynucleotide. Each selectable

!5   polynucleotide can thus have one or more end-selection markers, and they can be like or unlike end-selection markers. In a particular aspect, a plurality of end-selection markers can be located on one end of a polynucleotide and can have overlapping sequences with each other.

           It is important to emphasize that any number of enzymes, whether currently in

!0   existence or to be developed, can be serviceable in end-selection according to this invention. For example, in a particular aspect of this invention, a nicking enzyme (e.g. N. *Bst*NB I, which cleaves only one strand at 5'...GAGTCNNNN/N...3') can be used in conjunction

with a source of polynucleotide-ligating activity in order to achieve end-selection. According to this embodiment, a recognition site for N. *Bst*NB I – instead of a recognition site for topoisomerase I – should be incorporated into an end-selectable polynucleotide (whether end-selection is used for selection of a mutagenized progeny molecule or whether

5   end-selection is used apart from any mutagenesis procedure).

It is appreciated that the instantly disclosed end-selection approach using topoisomerase-based nicking and ligation has several advantages over previously available selection methods. In sum, this approach allows one to achieve direction cloning (including expression cloning). Specifically, this approach can be used for the achievement of: direct

10   ligation (i.e. without subjection to a classic restriction-purification-ligation reaction, that is susceptible to a multitude of potential problems from an initial restriction reaction to a ligation reaction dependent on the use of T4 DNA ligase); separation of progeny molecules from original template molecules (e.g. original template molecules lack topoisomerase I sites that not introduced until after mutagenesis), obviation of the need for size separation steps

15   (e.g. by gel chromatography or by other electrophoretic means or by the use of size-exclusion membranes), preservation of internal sequences (even when topoisomerase I sites are present), obviation of concerns about unsuccessful ligation reactions (e.g. dependent on the use of T4 DNA ligase, particularly in the presence of unwanted residual restriction enzyme activity), and facilitated expression cloning (including obviation of frame shift concerns).

20   Concerns about unwanted restriction enzyme-based cleavages – especially at internal restriction sites (or even at often unpredictable sites of unwanted star activity) in a working polynucleotide – that are potential sites of destruction of a working polynucleotide can also be obviated by the instantly disclosed end-selection approach using topoisomerase-based nicking and ligation.

ADDITIONAL SCREENING METHODS

Peptide Display Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by peptide display methods, wherein an associated polynucleotide encodes a displayed peptide which is screened for a phenotype (e.g., for affinity for a predetermined receptor (ligand).

An increasingly important aspect of bio-pharmaceutical drug development and molecular biology is the identification of peptide structures, including the primary amino acid sequences, of peptides or peptidomimetics that interact with biological macromolecules. one method of identifying peptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (e.g., a receptor), involves the screening of a large library or peptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the peptide.

In addition to direct chemical synthesis methods for generating peptide libraries, several recombinant DNA methods also have been reported. One type involves the display of a peptide sequence, antibody, or other protein on the surface of a bacteriophage particle or cell. Generally, in these methods each bacteriophage particle or cell serves as an individual library member displaying a single species of displayed peptide in addition to the natural bacteriophage or cell protein sequences. Each bacteriophage or cell contains the nucleotide sequence information encoding the particular displayed peptide sequence; thus, the displayed peptide sequence can be ascertained by nucleotide sequence determination of an isolated library member.

A well-known peptide display method involves the presentation of a peptide sequence on the surface of a filamentous bacteriophage, typically as a fusion with a bacteriophage coat protein. The bacteriophage library can be incubated with an immobilized, predetermined macromolecule or small molecule (e.g., a receptor) so that bacteriophage particles which present a peptide sequence that binds to the immobilized macromolecule can be differentially partitioned from those that do not present peptide sequences that bind to the predetermined macromolecule. The bacteriophage particles (i.e., library members) which are bound to the immobilized macromolecule are then recovered and replicated to amplify the selected

472

bacteriophage sub-population for a subsequent round of affinity enrichment and phage replication. After several rounds of affinity enrichment and phage replication, the bacteriophage library members that are thus selected are isolated and the nucleotide sequence encoding the displayed peptide sequence is determined, thereby identifying the sequence(s) of peptides that bind to the predetermined macromolecule (e.g., receptor). Such methods are further described in PCT patent publications WO 91/17271, WO 91/18980, WO 91/19818 and WO 93/08278.

The latter PCT publication describes a recombinant DNA method for the display of peptide ligands that involves the production of a library of fusion proteins with each fusion protein composed of a first polypeptide portion, typically comprising a variable sequence, that is available for potential binding to a predetermined macromolecule, and a second polypeptide portion that binds to DNA, such as the DNA vector encoding the individual fusion protein. When transformed host cells are cultured under conditions that allow for expression of the fusion protein, the fusion protein binds to the DNA vector encoding it. Upon lysis of the host cell, the fusion protein/vector DNA complexes can be screened against a predetermined macromolecule in much the same way as bacteriophage particles are screened in the phage-based display system, with the replication and sequencing of the DNA vectors in the selected fusion protein/vector DNA complexes serving as the basis for identification of the selected library peptide sequence(s).

Other systems for generating libraries of peptides and like polymers have aspects of both the recombinant and *in vitro* chemical synthesis methods. In these hybrid methods, cell-free enzymatic machinery is employed to accomplish the *in vitro* synthesis of the library members (i.e., peptides or polynucleotides). In one type of method, RNA molecules with the ability to bind a predetermined protein or a predetermined dye molecule were selected by alternate rounds of selection and PCR amplification (Tuerk and Gold, 1990; Ellington and Szostak, 1990). A similar technique was used to identify DNA sequences which bind a predetermined human transcription factor (Thiesen and Bach, 1990; Beaudry and Joyce, 1992; PCT patent publications WO 92/05258 and WO 92/14843). In a similar fashion, the technique of *in vitro* translation has been used to synthesize proteins of interest and has been proposed as a method for generating large libraries of peptides. These methods which rely upon *in vitro* translation, generally comprising stabilized polysome complexes, are described further in PCT patent publications WO 88/08453, WO 90/05785, WO 90/07003, WO

91/02076, WO 91/05058, and WO 92/02536. Applicants have described methods in which library members comprise a fusion protein having a first polypeptide portion with DNA binding activity and a second polypeptide portion having the library member unique peptide sequence; such methods are suitable for use in cell-free *in vitro* selection formats, among others.

The displayed peptide sequences can be of varying lengths, typically from 3-5000 amino acids long or longer, frequently from 5-100 amino acids long, and often from about 8-15 amino acids long. A library can comprise library members having varying lengths of displayed peptide sequence, or may comprise library members having a fixed length of displayed peptide sequence. Portions or all of the displayed peptide sequence(s) can be random, pseudorandom, defined set kernal, fixed, or the like. The present display methods include methods for *in vitro* and *in vivo* display of single-chain antibodies, such as nascent scFv on polysomes or scfv displayed on phage, which enable large-scale screening of scfv libraries having broad diversity of variable region sequences and binding specificities.

The present invention also provides random, pseudorandom, and defined sequence framework peptide libraries and methods for generating and screening those libraries to identify useful compounds (e.g., peptides, including single-chain antibodies) that bind to receptor molecules or epitopes of interest or gene products that modify peptides or RNA in a desired fashion. The random, pseudorandom, and defined sequence framework peptides are produced from libraries of peptide library members that comprise displayed peptides or displayed single-chain antibodies attached to a polynucleotide template from which the displayed peptide was synthesized. The mode of attachment may vary according to the specific embodiment of the invention selected, and can include encapsulation in a phage particle or incorporation in a cell.

A method of affinity enrichment allows a very large library of peptides and single-chain antibodies to be screened and the polynucleotide sequence encoding the desired peptide(s) or single-chain antibodies to be selected. The polynucleotide can then be isolated and shuffled to recombine combinatorially the amino acid sequence of the selected peptide(s) (or predetermined portions thereof) or single-chain antibodies (or just VHI, VLI or CDR portions thereof). Using these methods, one can identify a peptide or single-chain antibody as having a desired binding affinity for a molecule and can exploit the process of shuffling to converge rapidly to a desired high-affinity peptide or scfv. The peptide or antibody can then

be synthesized in bulk by conventional means for any suitable use (e.g., as a therapeutic or diagnostic agent).

A significant advantage of the present invention is that no prior information regarding an expected ligand structure is required to isolate peptide ligands or antibodies of interest. The peptide identified can have biological activity, which is meant to include at least specific binding affinity for a selected receptor molecule and, in some instances, will further include the ability to block the binding of other compounds, to stimulate or inhibit metabolic pathways, to act as a signal or messenger, to stimulate or inhibit cellular activity, and the like.

The present invention also provides a method for shuffling a pool of polynucleotide sequences selected by affinity screening a library of polysomes displaying nascent peptides (including single-chain antibodies) for library members which bind to a predetermined receptor (e.g., a mammalian proteinaceous receptor such as, for example, a peptidergic hormone receptor, a cell surface receptor, an intracellular protein which binds to other protein(s) to form intracellular protein complexes such as hetero-dimers and the like) or epitope (e.g., an immobilized protein, glycoprotein, oligosaccharide, and the like).

Polynucleotide sequences selected in a first selection round (typically by affinity selection for binding to a receptor (e.g., a ligand)) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral sequences (i.e., having insubstantial functional effect on binding), such as for example by back-crossing with a wild-type or naturally-occurring sequence substantially identical to a selected sequence to produce native-like functional peptides, which may be less immunogenic. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined receptor (ligand).

Prior to or concomitant with the shuffling of selected sequences, the sequences can be mutagenized. In one embodiment, selected library members are cloned in a prokaryotic vector (e.g., plasmid, phagemid, or bacteriophage) wherein a collection of individual colonies (or plaques) representing discrete library members are produced. Individual selected library

475

members can then be manipulated (e.g., by site-directed mutagenesis, cassette mutagenesis, chemical mutagenesis, PCR mutagenesis, and the like) to generate a collection of library members representing a kernal of sequence diversity based on the sequence of the selected library member. The sequence of an individual selected library member or pool can be

5   manipulated to incorporate random mutation, pseudorandom mutation, defined kernal mutation (i.e., comprising variant and invariant residue positions and/or comprising variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), codon-based mutation, and the like, either segmentally or over the entire length of the individual selected library member sequence. The mutagenized selected library members

10 are then shuffled by *in vitro* and/or *in vivo* recombinatorial shuffling as disclosed herein.

The invention also provides peptide libraries comprising a plurality of individual library members of the invention, wherein (1) each individual library member of said plurality comprises a sequence produced by shuffling of a pool of selected sequences, and (2) each individual library member comprises a variable peptide segment sequence or

15 single-chain antibody segment sequence which is distinct from the variable peptide segment sequences or single-chain antibody sequences of other individual library members in said plurality (although some library members may be present in more than one copy per library due to uneven amplification, stochastic probability, or the like).

The invention also provides a product-by-process, wherein selected polynucleotide

20 sequences having (or encoding a peptide having) a predetermined binding specificity are formed by the process of: (1) screening a displayed peptide or displayed single-chain antibody library against a predetermined receptor (e.g., ligand) or epitope (e.g., antigen macromolecule) and identifying and/or enriching library members which bind to the predetermined receptor or epitope to produce a pool of selected library members, (2)

25 shuffling by recombination the selected library members (or amplified or cloned copies thereof) which binds the predetermined epitope and has been thereby isolated and/or enriched from the library to generate a shuffled library, and (3) screening the shuffled library against the predetermined receptor (e.g., ligand) or epitope (e.g., antigen macromolecule) and identifying and/or enriching shuffled library members which bind to the predetermined

30 receptor or epitope to produce a pool of selected shuffled library members.

Antibody Display and Screening Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by antibody display methods, wherein an associated polynucleotide encodes a displayed antibody which is screened for a phenotype (e.g., for affinity for binding a predetermined antigen (ligand).

Various molecular genetic approaches have been devised to capture the vast immunological repertoire represented by the extremely large number of distinct variable regions which can be present in immunoglobulin chains. The naturally-occurring germ line immunoglobulin heavy chain locus is composed of separate tandem arrays of variable segment genes located upstream of a tandem array of diversity segment genes, which are themselves located upstream of a tandem array of joining (i) region genes, which are located upstream of the constant region genes. During B lymphocyte development, V-D-J rearrangement occurs wherein a heavy chain variable region gene (VH) is formed by rearrangement to form a fused D segment followed by rearrangement with a V segment to form a V-D-J joined product gene which, if productively rearranged, encodes a functional variable region (VH) of a heavy chain. Similarly, light chain loci rearrange one of several V segments with one of several J segments to form a gene encoding the variable region (VL) of a light chain.

The vast repertoire of variable regions possible in immunoglobulins derives in part from the numerous combinatorial possibilities of joining V and i segments (and, in the case of heavy chain loci, D segments) during rearrangement in B cell development. Additional sequence diversity in the heavy chain variable regions arises from non-uniform rearrangements of the D segments during V-D-J joining and from N region addition. Further, antigen-selection of specific B cell clones selects for higher affinity variants having non-germline mutations in one or both of the heavy and light chain variable regions; a phenomenon referred to as "affinity maturation" or "affinity sharpening". Typically, these "affinity sharpening" mutations cluster in specific areas of the variable region, most commonly in the complementarity-determining regions (CDRs).

In order to overcome many of the limitations in producing and identifying high-affinity immunoglobulins through antigen-stimulated β cell development (i.e., immunization), various prokaryotic expression systems have been developed that can be

477

manipulated to produce combinatorial antibody libraries which may be screened for high-affinity antibodies to specific antigens. Recent advances in the expression of antibodies in Escherichia coli and bacteriophage systems (see "alternative peptide display methods", *infra*) have raised the possibility that virtually any specificity can be obtained by either cloning antibody genes from characterized hybridomas or by de novo selection using antibody gene libraries (e.g., from Ig cDNA).

Combinatorial libraries of antibodies have been generated in bacteriophage lambda expression systems which may be screened as bacteriophage plaques or as colonies of lysogens (Huse et al, 1989); Caton and Koprowski, 1990; Mullinax et al, 1990; Persson et al, 1991). Various embodiments of bacteriophage antibody display libraries and lambda phage expression libraries have been described (Kang et al, 1991; Clackson et al, 1991; McCafferty et al, 1990; Burton et al, 1991; Hoogenboom et al, 1991; Chang et al, 1991; Breitling et al, 1991; Marks et al, 1991, p. 581; Barbas et al, 1992; Hawkins and Winter, 1992; Marks et al, 1992, p. 779; Marks et al, 1992, p. 16007; and Lowman et al, 1991; Lerner et al, 1992; all incorporated herein by reference). Typically, a bacteriophage antibody display library is screened with a receptor (e.g., polypeptide, carbohydrate, glycoprotein, nucleic acid) that is immobilized (e.g., by covalent linkage to a chromatography resin to enrich for reactive phage by affinity chromatography) and/or labeled (e.g., to screen plaque or colony lifts).

One particularly advantageous approach has been the use of so-called single-chain fragment variable (scfv) libraries (Marks et al, 1992, p. 779; Winter and Milstein, 1991; Clackson et al, 1991; Marks et al, 1991, p. 581; Chaudhary et al, 1990; Chiswell et al, 1992; McCafferty et al, 1990; and Huston et al, 1988). Various embodiments of scfv libraries displayed on bacteriophage coat proteins have been described.

Beginning in 1988, single-chain analogues of Fv fragments and their fusion proteins have been reliably generated by antibody engineering methods. The first step generally involves obtaining the genes encoding VH and VL domains with desired binding properties; these V genes may be isolated from a specific hybridoma cell line, selected from a combinatorial V-gene library, or made by V gene synthesis. The single-chain Fv is formed by connecting the component V genes with an oligonucleotide that encodes an appropriately designed linker peptide, such as (Gly-Gly-Gly-Gly-Ser)3 or equivalent linker peptide(s). The linker bridges the C-terminus of the first V region and N-terminus of the second, ordered as

either VH-linker-VL or VL-linker-VH' In principle, the scfv binding site can faithfully replicate both the affinity and specificity of its parent antibody combining site.

5 Thus, scfv fragments are comprised of VH and VL domains linked into a single polypeptide chain by a flexible linker peptide. After the scfv genes are assembled, they are cloned into a phagemid and expressed at the tip of the M13 phage (or similar filamentous bacteriophage) as fusion proteins with the bacteriophage PIII (gene 3) coat protein. Enriching for phage expressing an antibody of interest is accomplished by panning the recombinant phage displaying a population scfv for binding to a predetermined epitope (e.g., target antigen, receptor).

10 The linked polynucleotide of a library member provides the basis for replication of the library member after a screening or selection procedure, and also provides the basis for the determination, by nucleotide sequencing, of the identity of the displayed peptide sequence or VH and VL amino acid sequence. The displayed peptide (s) or single-chain antibody (e. g., scfv) and/or its VH and VL domains or their CDRs can be cloned and 15 expressed in a suitable expression system. Often polynucleotides encoding the isolated VH and VL domains will be ligated to polynucleotides encoding constant regions (CH and CL) to form polynucleotides encoding complete antibodies (e.g., chimeric or fully-human), antibody fragments, and the like. Often polynucleotides encoding the isolated CDRs will be grafted into polynucleotides encoding a suitable variable region framework (and optionally constant 20 regions) to form polynucleotides encoding complete antibodies (e.g., humanized or fully-human), antibody fragments, and the like. Antibodies can be used to isolate preparative quantities of the antigen by immunoaffinity chromatography. Various other uses of such antibodies are to diagnose and/or stage disease (e.g., neoplasia) and for therapeutic application to treat disease, such as for example: neoplasia, autoimmune disease, AIDS, 25 cardiovascular disease, infections, and the like.

Various methods have been reported for increasing the combinatorial diversity of a scfv library to broaden the repertoire of binding species (idiotype spectrum) The use of PCR has permitted the variable regions to be rapidly cloned either from a specific hybridoma source or as a gene library from non-immunized cells, affording combinatorial diversity in 30 the assortment of VH and VL cassettes which can be combined. Furthermore, the VH and VL cassettes can themselves be diversified, such as by random, pseudorandom, or directed mutagenesis. Typically, VH and VL cassettes are diversified in or near the

479

complementarity-determining regions (CDRS), often the third CDR, CDR3. Enzymatic inverse PCR mutagenesis has been shown to be a simple and reliable method for constructing relatively large libraries of scfv site-directed hybrids (Stemmer et al, 1993), as has error-prone PCR and chemical mutagenesis (Deng et al, 1994). Riechmann (Riechmann et al, 1993) showed semi-rational design of an antibody scfv fragment using site-directed randomization by degenerate oligonucleotide PCR and subsequent phage display of the resultant scfv hybrids. Barbas (Barbas et al, 1992) attempted to circumvent the problem of limited repertoire sizes resulting from using biased variable region sequences by randomizing the sequence in a synthetic CDR region of a human tetanus toxoid-binding Fab.

CDR randomization has the potential to create approximately 1×10^{20} CDRs for the heavy chain CDR3 alone, and a roughly similar number of variants of the heavy chain CDR1 and CDR2, and light chain CDR1-3 variants. Taken individually or together, the combination possibilities of CDR randomization of heavy and/or light chains requires generating a prohibitive number of bacteriophage clones to produce a clone library representing all possible combinations, the vast majority of which will be non-binding. Generation of such large numbers of primary transformants is not feasible with current transformation technology and bacteriophage display systems. For example, Barbas (Barbas et al, 1992) only generated 5×10^7 transformants, which represents only a tiny fraction of the potential diversity of a library of thoroughly randomized CDRS.

Despite these substantial limitations, bacteriophage. display of scfv have already yielded a variety of useful antibodies and antibody fusion proteins. A bispecific single chain antibody has been shown to mediate efficient tumor cell lysis (Gruber et al, 1994). Intracellular expression of an anti-Rev scfv has been shown to inhibit HIV-1 virus replication *in vitro* (Duan et al, 1994), and intracellular expression of an anti-p21rar, scfv has been shown to inhibit meiotic maturation of <u>Xenopus</u> oocytes (Biocca et al, 1993). Recombinant scfv which can be used to diagnose HIV infection have also been reported, demonstrating the diagnostic utility of scfv (Lilley et al, 1994). Fusion proteins wherein an scFv is linked to a second polypeptide, such as a toxin or fibrinolytic activator protein, have also been reported (Holvost et al, 1992; Nicholls et al, 1993).

If it were possible to generate scfv libraries having broader antibody diversity and overcoming many of the limitations of conventional CDR mutagenesis and randomization methods which can cover only a very tiny fraction of the potential sequence combinations,

the number and quality of scfv antibodies suitable for therapeutic and diagnostic use could be vastly improved. To address this, the *in vitro* and *in vivo* shuffling methods of the invention are used to recombine CDRs which have been obtained (typically via PCR amplification or cloning) from nucleic acids obtained from selected displayed antibodies. Such displayed

5   antibodies can be displayed on cells, on bacteriophage particles, on polysomes, or any suitable antibody display system wherein the antibody is associated with its encoding nucleic acid(s). In a variation, the CDRs are initially obtained from mRNA (or cDNA) from antibody-producing cells (e.g., plasma cells/splenocytes from an immunized wild-type mouse, a human, or a transgenic mouse capable of making a human antibody as in WO

10   92/03918, WO 93/12227, and WO 94/25585), including hybridomas derived therefrom.

Polynucleotide sequences selected in a first selection round (typically by affinity selection for displayed antibody binding to an antigen (e.g., a ligand) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination, especially shuffling of CDRs (typically shuffling heavy chain CDRs with other heavy chain

15   CDRs and light chain CDRs with other light chain CDRs) to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are expressed in a selection format as a displayed antibody and subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or

20   subjected to one or more additional rounds of shuffling and subsequent selection until an antibody of the desired binding affinity is obtained. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral antibody framework sequences (i.e., having insubstantial functional effect on antigen binding), such as for example by back-crossing with a human variable region framework to produce human-like sequence

25   antibodies. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined antigen.

Alternatively, or in combination with the noted variations, the valency of the target epitope may be varied to control the average binding affinity of selected scfv library members. The target epitope can be bound to a surface or substrate at varying densities, such

30   as by including a competitor epitope, by dilution, or by other method known to those in the art. A high density (valency) of predetermined epitope can be used to enrich for scfv library

members which have relatively low affinity, whereas a low density (valency) can preferentially enrich for higher affinity scfv library members.

For generating diverse variable segments, a collection of synthetic oligonucleotides encoding random, pseudorandom, or a defined sequence kernal set of peptide sequences can be inserted by ligation into a predetermined site (e.g., a CDR). Similarly, the sequence diversity of one or more CDRs of the single-chain antibody cassette(s) can be expanded by mutating the CDR(s) with site-directed mutagenesis, CDR-replacement, and the like. The resultant DNA molecules can be propagated in a host for cloning and amplification prior to shuffling, or can be used directly (i.e., may avoid loss of diversity which may occur upon propagation in a host cell) and the selected library members subsequently shuffled.

Displayed peptide/polynucleotide complexes (library members) which encode a variable segment peptide sequence of interest or a single-chain antibody of interest are selected from the library by an affinity enrichment technique. This is accomplished by means of a immobilized macromolecule or epitope specific for the peptide sequence of interest, such as a receptor, other macromolecule, or other epitope species. Repeating the affinity selection procedure provides an enrichment of library members encoding the desired sequences, which may then be isolated for pooling and shuffling, for sequencing, and/or for further propagation and affinity enrichment.

The library members without the desired specificity are removed by washing. The degree and stringency of washing required will be determined for each peptide sequence or single-chain antibody of interest and the immobilized predetermined macromolecule or epitope. A certain degree of control can be exerted over the binding characteristics of the nascent peptide/DNA complexes recovered by adjusting the conditions of the binding incubation and the subsequent washing. The temperature, pH, ionic strength, divalent cations concentration, and the volume and duration of the washing will select for nascent peptide/DNA complexes within particular ranges of affinity for the immobilized macromolecule. Selection based on slow dissociation rate, which is usually predictive of high affinity, is often the most practical route. This may be done either by continued incubation in the presence of a saturating amount of free predetermined macromolecule, or by increasing the volume, number, and length of the washes. In each case, the rebinding of dissociated nascent peptide/DNA or peptide/RNA complex is prevented, and with increasing

time, nascent peptide/DNA or peptide/RNA complexes of higher and higher affinity are recovered.

Additional modifications of the binding and washing procedures may be applied to find peptides with special characteristics. The affinities of some peptides are dependent on ionic strength or cation concentration. This is a useful characteristic for peptides that will be used in affinity purification of various proteins when gentle conditions for removing the protein from the peptides are required.

One variation involves the use of multiple binding targets (multiple epitope species, multiple receptor species), such that a scfv library can be simultaneously screened for a multiplicity of scfv which have different binding specificities. Given that the size of a scfv library often limits the diversity of potential scfv sequences, it is typically desirable to us scfv libraries of as large a size as possible. The time and economic considerations of generating a number of very large polysome scFv-display libraries can become prohibitive. To avoid this substantial problem, multiple predetermined epitope species (receptor species) can be concomitantly screened in a single library, or sequential screening against a number of epitope species can be used. In one variation, multiple target epitope species, each encoded on a separate bead (or subset of beads), can be mixed and incubated with a polysome-display scfv library under suitable binding conditions. The collection of beads, comprising multiple epitope species, can then be used to isolate, by affinity selection, scfv library members. Generally, subsequent affinity screening rounds can include the same mixture of beads, subsets thereof, or beads containing only one or two individual epitope species. This approach affords efficient screening, and is compatible with laboratory automation, batch processing, and high throughput screening methods.

A variety of techniques can be used in the present invention to diversify a peptide library or single-chain antibody library, or to diversify, prior to or concomitant with shuffling, around variable segment peptides found in early rounds of panning to have sufficient binding activity to the predetermined macromolecule or epitope. In one approach, the positive selected peptide/polynucleotide complexes (those identified in an early round of affinity enrichment) are sequenced to determine the identity of the active peptides. Oligonucleotides are then synthesized based on these active peptide sequences, employing a low level of all bases incorporated at each step to produce slight variations of the primary oligonucleotide sequences. This mixture of (slightly) degenerate oligonucleotides is then

cloned into the variable segment sequences at the appropriate locations. This method produces systematic, controlled variations of the starting peptide sequences, which can then be shuffled. It requires, however, that individual positive nascent peptide/polynucleotide complexes be sequenced before mutagenesis, and thus is useful for expanding the diversity of small numbers of recovered complexes and selecting variants having higher binding affinity and/or higher binding specificity. In a variation, mutagenic PCR amplification of positive selected peptide/polynucleotide complexes (especially of the variable region sequences, the amplification products of which are shuffled *in vitro* and/or *in vivo* and one or more additional rounds of screening is done prior to sequencing. The same general approach can be employed with single-chain antibodies in order to expand the diversity and enhance the binding affinity/specificity, typically by diversifying CDRs or adjacent framework regions prior to or concomitant with shuffling. If desired, shuffling reactions can be spiked with mutagenic oligonucleotides capable of *in vitro* recombination with the selected library members can be included. Thus, mixtures of synthetic oligonucleotides and PCR produced polynucleotides (synthesized by error-prone or high-fidelity methods) can be added to the *in vitro* shuffling mix and be incorporated into resulting shuffled library members (shufflants).

The present invention of shuffling enables the generation of a vast library of CDR-variant single-chain antibodies. One way to generate such antibodies is to insert synthetic CDRs into the single-chain antibody and/or CDR randomization prior to or concomitant with shuffling. The sequences of the synthetic CDR cassettes are selected by referring to known sequence data of human CDR and are selected in the discretion of the practitioner according to the following guidelines: synthetic CDRs will have at least 40 percent positional sequence identity to known CDR sequences, and preferably will have at least 50 to 70 percent positional sequence identity to known CDR sequences. For example, a collection of synthetic CDR sequences can be generated by synthesizing a collection of oligonucleotide sequences on the basis of naturally-occurring human CDR sequences listed in Kabat (Kabat et al, 1991); the pool (s) of synthetic CDR sequences are calculated to encode CDR peptide sequences having at least 40 percent sequence identity to at least one known naturally-occurring human CDR sequence. Alternatively, a collection of naturally-occurring CDR sequences may be compared to generate consensus sequences so that amino acids used at a residue position frequently (i.e., in at least 5 percent of known CDR sequences) are incorporated into the synthetic CDRs at the corresponding position(s).

Typically, several (e.g., 3 to about 50) known CDR sequences are compared and observed natural sequence variations between the known CDRs are tabulated, and a collection of oligonucleotides encoding CDR peptide sequences encompassing all or most permutations of the observed natural sequence variations is synthesized. For example but not for limitation, if a collection of human VH CDR sequences have carboxy-terminal amino acids which are either Tyr, Val, Phe, or Asp, then the pool(s) of synthetic CDR oligonucleotide sequences are designed to allow the carboxy-terminal CDR residue to be any of these amino acids. In some embodiments, residues other than those which naturally-occur at a residue position in the collection of CDR sequences are incorporated: conservative amino acid substitutions are frequently incorporated and up to 5 residue positions may be varied to incorporate non-conservative amino acid substitutions as compared to known naturally-occurring CDR sequences. Such CDR sequences can be used in primary library members (prior to first round screening) and/or can be used to spike *in vitro* shuffling reactions of selected library member sequences. Construction of such pools of defined and/or degenerate sequences will be readily accomplished by those of ordinary skill in the art.

The collection of synthetic CDR sequences comprises at least one member that is not known to be a naturally-occurring CDR sequence. It is within the discretion of the practitioner to include or not include a portion of random or pseudorandom sequence corresponding to N region addition in the heavy chain CDR; the N region sequence ranges from 1 nucleotide to about 4 nucleotides occurring at V-D and D-J junctions. A collection of synthetic heavy chain CDR sequences comprises at least about 100 unique CDR sequences, typically at least about 1,000 unique CDR sequences, preferably at least about 10,000 unique CDR sequences, frequently more than 50,000 unique CDR sequences; however, usually not more than about 1×10^6 unique CDR sequences are included in the collection, although occasionally 1×10^7 to 1×10^8 unique CDR sequences are present, especially if conservative amino acid substitutions are permitted at positions where the conservative amino acid substituent is not present or is rare (i.e., less than 0.1 percent) in that position in naturally-occurring human CDRS. In general, the number of unique CDR sequences included in a library should not exceed the expected number of primary transformants in the library by more than a factor of 10. Such single-chain antibodies generally bind of about at least 1×10 m-, preferably with an affinity of about at least 5×10^7 M-1, more preferably with an affinity of at least 1×10^8 M-1 to 1×10^9 M-1 or more, sometimes up to 1×10^{10} M-1 or more.

Frequently, the predetermined antigen is a human protein, such as for example a human cell surface antigen (e. g., CD4, CD8, IL-2 receptor, EGF receptor, PDGF receptor), other human biological macromolecule (e.g., thrombomodulin, protein C, carbohydrate antigen, sialyl Lewis antigen, Lselectin), or nonhuman disease associated macromolecule (e.g., bacterial

5   LPS, virion capsid protein or envelope glycoprotein) and the like.

High affinity single-chain antibodies of the desired specificity can be engineered and expressed in a variety of systems. For example, scfv have been produced in plants (Firek et al, 1993) and can be readily made in prokaryotic systems (Owens and Young, 1994; Johnson and Bird, 1991). Furthermore, the single-chain antibodies can be used as a basis for

10   constructing whole antibodies or various fragments thereof (Kettleborough et al, 1994). The variable region encoding sequence may be isolated (e.g., by PCR amplification or subcloning) and spliced to a sequence encoding a desired human constant region to encode a human sequence antibody more suitable for human therapeutic uses where immunogenicity is preferably minimized. The polynucleotide(s) having the resultant fully human encoding

15   sequence(s) can be expressed in a host cell (e.g., from an expression vector in a mammalian cell) and purified for pharmaceutical formulation.

The DNA expression constructs will typically include an expression control DNA sequence operably linked to the coding sequences, including naturally-associated or heterologous promoter regions. Preferably, the expression control sequences will be

20   eukaryotic promoter systems in vectors capable of transforming or transfecting eukaryotic host cells. Once the vector has been incorporated into the appropriate host, the host is maintained under conditions suitable for high level expression of the nucleotide sequences, and the collection and purification of the mutant' "engineered" antibodies.

As stated previously, the DNA sequences will be expressed in hosts after the

25   sequences have been operably linked to an expression control sequence (i.e., positioned to ensure the transcription and translation of the structural gene). These expression vectors are typically replicable in the host organisms either as episomes or as an integral part of the host chromosomal DNA. Commonly, expression vectors will contain selection markers, e.g., tetracycline or neomycin, to permit detection of those cells transformed with the desired

30   DNA sequences (see, e.g., USPN 4,704,362, which is incorporated herein by reference).

In addition to eukaryotic microorganisms such as yeast, mammalian tissue cell culture may also be used to produce the polypeptides of the present invention (see Winnacker, 1987),

which is incorporated herein by reference). Eukaryotic cells are actually preferred, because a number of suitable host cell lines capable of secreting intact immunoglobulins have been developed in the art, and include the CHO cell lines, various COS cell lines, HeLa cells, and myeloma cell lines, but preferably transformed Bcells or hybridomas. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter, an enhancer (Queen et al, 1986), and necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites, and transcriptional terminator sequences. Preferred expression control sequences are promoters derived from immunoglobulin genes, cytomegalovirus, SV40, Adenovirus, Bovine Papilloma Virus, and the like.

Eukaryotic DNA transcription can be increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting sequences of between 10 to 300 bp that increase transcription by a promoter. Enhancers can effectively increase transcription when either 51 or 31 to the transcription unit. They are also effective if located within an intron or within the coding sequence itself. Typically, viral enhancers are used, including SV40 enhancers, cytomegalovirus enhancers, polyoma enhancers, and adenovirus enhancers. Enhancer sequences from mammalian systems are also commonly used, such as the mouse immunoglobulin heavy chain enhancer.

Mammalian expression vector systems will also typically include a selectable marker gene. Examples of suitable markers include, the dihydrofolate reductase gene (DHFR), the thymidine kinase gene (TK), or prokaryotic genes conferring drug resistance. The first two marker genes prefer the use of mutant cell lines that lack the ability to grow without the addition of thymidine to the growth medium. Transformed cells can then be identified by their ability to grow on non-supplemented media. Examples of prokaryotic drug resistance genes useful as markers include genes conferring resistance to G418, mycophenolic acid and hygromycin.

The vectors containing the DNA segments of interest can be transferred into the host cell by well-known methods, depending on the type of cellular host. For example, calcium chloride transfection is commonly utilized for prokaryotic cells, whereas calcium phosphate treatment. lipofection, or electroporation may be used for other cellular hosts. Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion,

liposomes, electroporation, and micro-injection (see, generally, Sambrook et al, 1982 and 1989).

Once expressed, the antibodies, individual mutated immunoglobulin chains, mutated antibody fragments, and other immunoglobulin polypeptides of the invention can be purified according to standard procedures of the art, including ammonium sulfate precipitation, fraction column chromatography, gel electrophoresis and the like (see, generally, Scopes, 1982). Once purified, partially or to homogeneity as desired, the polypeptides may then be used therapeutically or in developing and performing assay procedures, immunofluorescent stainings, and the like (see, generally, Lefkovits and Pernis, 1979 and 1981; Lefkovits, 1997).

The antibodies generated by the method of the present invention can be used for diagnosis and therapy. By way of illustration and not limitation, they can be used to treat cancer, autoimmune diseases, or viral infections. For treatment of cancer, the antibodies will typically bind to an antigen expressed preferentially on cancer cells, such as erbB-2, CEA, CD33, and many other antigens and binding members well known to those skilled in the art.

Two-Hybrid Based Screening Assays

Shuffling can also be used to recombinatorially diversify a pool of selected library members obtained by screening a two-hybrid screening system to identify library members which bind a predetermined polypeptide sequence. The selected library members are pooled and shuffled by *in vitro* and/or *in vivo* recombination. The shuffled pool can then be screened in a yeast two hybrid system to select library members which bind said predetermined polypeptide sequence (e. g., and SH2 domain) or which bind an alternate predetermined polypeptide sequence (e.g., an SH2 domain from another protein species).

An approach to identifying polypeptide sequences which bind to a predetermined polypeptide sequence has been to use a so-called "two-hybrid" system wherein the predetermined polypeptide sequence is present in a fusion protein (Chien et al, 1991). This approach identifies protein-protein interactions *in vivo* through reconstitution of a transcriptional activator (Fields and Song, 1989), the yeast Gal4 transcription protein. Typically, the method is based on the properties of the yeast Gal4 protein, which consists of separable domains responsible for DNA-binding and transcriptional activation. Polynucleotides encoding two hybrid proteins, one consisting of the yeast Gal4 DNA-binding domain fused to a polypeptide sequence of a known protein and the other consisting of the Gal4 activation domain fused to a polypeptide sequence of a second protein, are constructed

488

and introduced into a yeast host cell. Intermolecular binding between the two fusion proteins reconstitutes the Gal4 DNA-binding domain with the Gal4 activation domain, which leads to the transcriptional activation of a reporter gene (e.g., *lacz, HIS3*) which is operably linked to a Gal4 binding site. Typically, the two-hybrid method is used to identify novel polypeptide

5   sequences which interact with a known protein (Silver and Hunt, 1993; Durfee et al, 1993; Yang et al, 1992; Luban et al, 1993; Hardy et al, 1992; Bartel et al, 1993; and Vojtek et al, 1993). However, variations of the two-hybrid method have been used to identify mutations of a known protein that affect its binding to a second known protein (Li and Fields, 1993; Lalo et al, 1993; Jackson et al, 1993; and Madura et al, 1993). Two-hybrid systems have also

10   been used to identify interacting structural domains of two known proteins (Bardwell et al, 1993; Chakrabarty et al, 1992; Staudinger et al, 1993; and Milne and Weaver 1993) or domains responsible for oligomerization of a single protein (Iwabuchi et al, 1993; Bogerd et al, 1993). Variations of two-hybrid systems have been used to study the *in vivo* activity of a proteolytic enzyme (Dasmahapatra et al, 1992). Alternatively, an E. coli/BCCP interactive

15   screening system (Germino et al, 1993; Guarente, 1993) can be used to identify interacting protein sequences (i.e., protein sequences which heterodimerize or form higher order heteromultimers). Sequences selected by a two-hybrid system can be pooled and shuffled and introduced into a two-hybrid system for one or more subsequent rounds of screening to identify polypeptide sequences which bind to the hybrid containing the predetermined

20   binding sequence. The sequences thus identified can be compared to identify consensus sequence(s) and consensus sequence kernals.

In general, standard techniques of recombination DNA technology are described in various publications (e.g. Sambrook et al, 1989; Ausubel et al, 1987; and Berger and Kimmel, 1987; each of which is incorporated herein in its entirety by reference.

25   Polynucleotide modifying enzymes were used according to the manufacturer's recommendations. Oligonucleotides were synthesized on an Applied Biosystems Inc. Model 394 DNA synthesizer using ABI chemicals. If desired, PCR amplimers for amplifying a predetermined DNA sequence may be selected at the discretion of the practitioner.

One microgram samples of template DNA are obtained and treated with U.V. light to

30   cause the formation of dimers, including TT dimers, particularly purine dimers. U.V. exposure is limited so that only a few photoproducts are generated per gene on the template

489

DNA sample. Multiple samples are treated with U.V. light for varying periods of time to obtain template DNA samples with varying numbers of dimers from U.V. exposure.

A random priming kit which utilizes a non-proofreading polymease (for example, Prime-It II Random Primer Labeling kit by Stratagene Cloning Systems) is utilized to generate different size polynucleotides by priming at random sites on templates which are prepared by U.V. light (as described above) and extending along the templates. The priming protocols such as described in the Prime-It II Random Primer Labeling kit may be utilized to extend the primers. The dimers formed by U.V. exposure serve as a roadblock for the extension by the non-proofreading polymerase. Thus, a pool of random size polynucleotides is present after extension with the random primers is finished.

The present invention is further directed to a method for generating a selected mutant polynucleotide sequence (or a population of selected polynucleotide sequences) typically in the form of amplified and/or cloned polynucleotides, whereby the selected polynucleotide sequences(s) possess at least one desired phenotypic characteristic (e.g., encodes a polypeptide, promotes transcription of linked polynucleotides, binds a protein, and the like) which can be selected for. One method for identifying hybrid polypeptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (e.g., a receptor), involves the screening of a large library of polypeptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the polypeptide.

In one embodiment, the present invention provides a method for generating libraries of displayed polypeptides or displayed antibodies suitable for affinity interaction screening or phenotypic screening. The method comprises (1) obtaining a first plurality of selected library members comprising a displayed polypeptide or displayed antibody and an associated polynucleotide encoding said displayed polypeptide or displayed antibody, and obtaining said associated polynucleotides or copies thereof wherein said associated polynucleotides comprise a region of substantially identical sequences, optimally introducing mutations into said polynucleotides or copies, (2) pooling the polynucleotides or copies, (3) producing smaller or shorter polynucleotides by interrupting a random or particularized priming and synthesis process or an amplification process, and (4) performing amplification, preferably PCR amplification, and optionally mutagenesis to homologously recombine the newly synthesized polynucleotides.

490

In one aspect, the invention provides a process for producing hybrid polynucleotides which express a useful hybrid polypeptide by a series of steps comprising:

(a)   producing polynucleotides by interrupting a polynucleotide amplification or synthesis process with a means for blocking or interrupting the amplification or synthesis process and thus providing a plurality of smaller or shorter polynucleotides due to the replication of the polynucleotide being in various stages of completion;

(b)   adding to the resultant population of single- or double-stranded polynucleotides one or more single- or double-stranded oligonucleotides, wherein said added oligonucleotides comprise an area of identity in an area of heterology to one or more of the single- or double-stranded polynucleotides of the population;

(c)   denaturing the resulting single- or double-stranded oligonucleotides to produce a mixture of single-stranded polynucleotides, optionally separating the shorter or smaller polynucleotides into pools of polynucleotides having various lengths and further optionally subjecting said polynucleotides to a PCR procedure to amplify one or more oligonucleotides comprised by at least one of said polynucleotide pools;

(d)   incubating a plurality of said polynucleotides or at least one pool of said polynucleotides with a polymerase under conditions which result in annealing of said single-stranded polynucleotides at regions of identity between the single-stranded polynucleotides and thus forming of a mutagenized double-stranded polynucleotide chain;

(e)   optionally repeating steps (c) and (d);

(f)   expressing at least one hybrid polypeptide from said polynucleotide chain, or chains; and

(g)   screening said at least one hybrid polypeptide for a useful activity.

In a preferred aspect of the invention, the means for blocking or interrupting the amplification or synthesis process is by utilization of uv light, DNA adducts, DNA binding proteins.

In one embodiment of the invention, the DNA adducts, or polynucleotides comprising the DNA adducts, are removed from the polynucleotides or polynucleotide pool, such as by a process including heating the solution comprising the DNA fragments prior to further processing.

491

It will be readily apparent to one skilled in the art that various substitutions and modifications may be made to the invention disclosed herein without departing from the scope and spirit of the invention. It is understood that the examples and aspects described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

EXAMPLES

The following example is offered to illustrate, but not to limit the claimed invention.

Example 1
Generation of Random Size Polynucleotides Using U.V. Induced Photoproducts

One microgram samples of template DNA are obtained and treated with U.V. light to cause the formation of dimers, including TT dimers, particularly purine dimers. U.V. exposure is limited so that only a few photoproducts are generated per gene on the template DNA sample. Multiple samples are treated with U.V. light for varying periods of time to obtain template DNA samples with varying numbers of dimers from U.V. exposure.

A random priming kit which utilizes a non-proofreading polymerase (for example, Prime-It II Random Primer Labeling kit by Stratagene Cloning Systems) is utilized to generate different size polynucleotides by priming at random sites on templates which are prepared by U.V. light (as described above) and extending along the templates. The priming protocols such as described in the Prime-It II Random Primer Labeling kit may be utilized to extend the primers. The dimers formed by U.V. exposure serve as a roadblock for the extension by the non-proofreading polymerase. Thus, a pool of random size polynucleotides is present after extension with the random primers is finished.

Example 2
Isolation of Random Size Polynucleotides

Polynucleotides of interest which are generated according to Example 1 are gel isolated on a 1.5% agarose gel. Polynucleotides in the 100-300 bp range are cut out of the gel and 3 volumes of 6 M NaI is added to the gel slice. The mixture is incubated at 50 °C for 10 minutes and 10 µl of glass milk (Bio 101) is added. The mixture is spun for 1 minute and

the supernatant is decanted. The pellet is washed with 500 µl of Column Wash (Column Wash is 50% ethanol, 10mM Tris-HCl pH 7.5, 100 mM NaCl and 2.5 mM EDTA) and spin for 1 minute, after which the supernatant is decanted. The washing, spinning and decanting steps are then repeated. The glass milk pellet is resuspended in 20µl of H_2O and spun for 1 minute. DNA remains in the aqueous phase.

Example 3
Shuffling of Isolated Random Size 100-300bp Polynucleotides

The 100-300 bp polynucleotides obtained in Example 2 are recombined in an annealing mixture (0.2 mM each dNTP, 2.2 mM $MgCl_2$, 50 mM KCl, 10 mM Tris-HCl ph 8.8, 0.1% Triton X-100, 0.3 µ; Taq DNA polymerase, 50 µl total volume) without adding primers. A Robocycler by Stratagene was used for the annealing step with the following program: 95 °C for 30 seconds, 25-50 cycles of [95 °C for 30 seconds, 50 - 60 °C (preferably 58 °C) for 30 seconds, and 72 °C for 30 seconds] and 5 minutes at 72 °C. Thus, the 100-300 bp polynucleotides combine to yield double-stranded polynucleotides having a longer sequence. After separating out the reassembled double-stranded polynucleotides and denaturing them to form single stranded polynucleotides, the cycling is optionally again repeated with some samples utilizing the single strands as template and primer DNA and other samples utilizing random primers in addition to the single strands.

Example 4
Screening of Polypeptides from Shuffled Polynucleotides

The polynucleotides of Example 3 are separated and polypeptides are expressed therefrom. The original template DNA is utilized as a comparative control by obtaining comparative polypeptides therefrom. The polypeptides obtained from the shuffled polynucleotides of Example 3 are screened for the activity of the polypeptides obtained from the original template and compared with the activity levels of the control. The shuffled polynucleotides coding for interesting polypeptides discovered during screening are compared further for secondary desirable traits. Some shuffled polynucleotides corresponding to less interesting screened polypeptides are subjected to reshuffling.

Example 5

Directed Evolution an Enzyme by Saturation Mutagenesis

Site-Saturation Mutagenesis: To accomplish site-saturation mutagenesis every residue (316) of a dehalogenase enzyme was converted into all 20 amino acids by site directed mutagenesis using 32-fold degenerate oligonucleotide primers, as follows:

1. A culture of the dehalogenase expression construct was grown and a preparation of the plasmid was made

2. Primers were made to randomize each codon - they have the common structure $X_{20}NN(G/T)X_{20}$

3. A reaction mix of 25 ul was prepared containing ~50 ng of plasmid template, 125 ng of each primer, 1X native Pfu buffer, 200 uM each dNTP and 2.5 U native Pfu DNA polymerase

4. The reaction was cycled in a Robo96 Gradient Cycler as follows:
    Initial denaturation at 95°C for 1 min
    20 cycles of 95°C for 45 sec, 53°C for 1 min and 72°C for 11 min
    Final elongation step of 72°C for 10 min

5. The reaction mix was digested with 10 U of DpnI at 37°C for 1 hour to digest the methylated template DNA

6. Two ul of the reaction mix were used to transform 50 ul of XL1-Blue MRF' cells and the entire transformation mix was plated on a large LB-Amp-Met plate yielding 200-1000 colonies

7. Individual colonies were toothpicked into the wells of 96-well microtiter plates containing LB-Amp-IPTG and grown overnight

8. The clones on these plates were assayed the following day

Screening: Approximately 200 clones of mutants for each position were grown in liquid media (384 well microtiter plates) and screened as follows:

1. Overnight cultures in 384-well plates were centrifuged and the media removed. To each well was added 0.06 mL 1 mM Tris/SO_4^{2-} pH 7.8.

2. Made 2 assay plates from each parent growth plate consisting of 0.02 mL cell suspension.

494

3.   One assay plate was placed at room temperature and the other at elevated temperature (initial screen used 55°C) for a period of time (initially 30 minutes).

4.   After the prescribed time 0.08 mL room temperature substrate (TCP saturated 1 mM Tris/SO_4^{2-} pH 7.8 with 1.5 mM NaN_3 and 0.1 mM bromothymol blue) was added to each well.

5.   Measurements at 620 nm were taken at various time points to generate a progress curve for each well.

6.   Data were analyzed and the kinetics of the cells heated to those not heated were compared.   Each plate contained 1-2 columns (24 wells) of unmutated 20F12 controls.

7.   Wells that appeared to have improved stability were re-grown and tested under the same conditions.

Following this procedure nine single site mutations appeared to confer increased thermal stability on the enzyme. Sequence analysis was performed to determine of the exact amino acid changes at each position that were specifically responsible for the improvement. In sum, the improvement was conferred at 7 sites by one amino acid change alone, at an eighth site by each of two amino acid changes, and at a ninth site by each of three amino acid changes. Several mutants were then made each having a plurality of these nine beneficial site mutations in combination; of these two mutants proved superior to all the other mutants, including those with single point mutations.

Example 6
Direct expression cloning using end-selection

An esterase gene was amplified using 5'phosphorylated primers in a standard PCR reaction (10 ng template; PCR conditions: 3' 94 C; [1' 94 C; 1' 50 C; 1'30" 68 C] x 30; 10' 68 C.

Forward   Primer   =   9511TopF
(CTAGAAGGGAGGAGAATTACATGAAGCGGCTTTTAGCCC)
Reverse Primer = 9511TopR (AGCTAAGGGTCAAGGCCGCACCCGAGG)
The resulting PCR product (ca.1000 bp) was gel purified and quantified.

A vector for expression cloning, pASK3 (Institut fuer Bioanalytik, Goettingen, Germany), was cut with *Xba* I and *Bgl* II and dephosphorylated with CIP.

0.5 pmoles Vaccina Topoisomerase I (Invitrogen, Carlsbad, CA) was added to 60 ng
5   (ca. 0.1 pmole) purified PCR product for 5' 37 C in buffer NEB I (New England Biolabs, Beverly, MA) in 5 µl total volume.

The topogated PCR product was cloned into the vector pASK3 (5 µl, ca. 200 ng in NEB I) for 5' at room temperature.

This mixture was dialyzed against H$_2$O for 30'.

10   2 µl were used for electroporation of DH10B cells (Gibco BRL, Gaithersburg, MD).

Efficiency: Based on the actual clone numbers this method can produce 2 x 10^6 clones per µg vector. All tested recombinants showed esterase activity after induction with anhydrotetracycline.

15   **Example 7**
                  Dehalogenase Thermal Stability

This invention provides that a desirable property to be generated by directed evolution is exemplified in a limiting fashion by an improved residual activity (e.g. an enzymatic activity, an immunoreactivity, an antibiotic acivity, etc.) of a molecule upon
20   subjection to altered environment, including what may be considered a harsh enviroment, for a specified time. Such a harsh environment may comprise any combination of the following (iteratively or not, and in any order or permutation): an elevated temperature (including a temperature that may cause denaturation of a working enzyme), a decreased temperature, an elevated salinity, a decreased salinity, an elevated pH, a decreased pH, an elevated pressure, a
25   decreassed pressure, and an change in exposure to a radiation source (including uv radiation, visible light, as well as the entire electromagnetic spectrum).

The following example shows an application of directed evolution to evolve the ability of an enzyme to regain &/or retain activity upon exposure to an elevated temperature.

Every residue (316) of a dehalogenase enzyme was converted into all 20 amino acids
30   by site directed mutagenesis using 32-fold degenerate oligonucleotide primers. These mutations were introduced into the already rate-improved variant Dhla 20F12.

Approximately 200 clones of each position were grown in liquid media (384 well microtiter plates) to be screened. The screening procedure was as follows:

1. Overnight cultures in 384-well plates were centrifuged and the media removed. To each well was added 0.06 mL 1 mM Tris/SO$_4^{2-}$ pH 7.8.

5

2. The robot made 2 assay plates from each parent growth plate consisting of 0.02 mL cell suspension.

3. One assay plate was placed at room temperature and the other at elevated temperature (initial screen used 55°C) for a period of time (initially 30 minutes).

4. After the prescribed time 0.08 mL room temperature substrate (TCP saturated 1 mM

10

Tris/SO$_4^{2-}$ pH 7.8 with 1.5 mM NaN$_3$ and 0.1 mM bromothymol blue) was added to each well. TCP = trichloropropane.

5. Measurements at 620 nm were taken at various time points to generate a progress curve for each well.

6. Data were analyzed and the kinetics of the cells heated to those not heated were

15

compared. Each plate contained 1-2 columns (24 wells) of un-mutated 20F12 controls.

7. Wells that appeared to have improved stability were regrown and tested under the same conditions.

Following this procedure nine single site mutations appeared to confer increased

20

thermal stability on Dhla-20F12. Sequence analysis showed that the following changes were beneficial:

D89G

F91S

T159L

25

G189Q, G189V

I220L

N238T

W251Y

P302A, P302L, P302S, P302K

30

P302R/S306R

Only two sites (189 and 302) had more than one substitution. The first 5 on the list were combined (using G189Q) into a single gene (this mutant is referred to as "Dhla5"). All changes but S306R were incorporated into another variant referred to as Dhla8.

Thermal stability was assessed by incubating the enzyme at the elevated temperature (55°C and 80°C) for some period of time and activity assay at 30°C. Initial rates were plotted vs. time at the higher temperature. The enzyme was in 50 mM Tris/SO$_4$ pH 7.8 for both the incubation and the assay. Product (Cl$^-$) was detected by a standard method using Fe(NO$_3$)$_3$ and HgSCN. Dhla 20F12 was used as the *de facto* wild type. The apparent half-life (T$_{1/2}$) was calculated by fitting the data to an exponential decay function.

Example 8:
Saturation mutagenesis

The following example describes manipulation of three related parental nucleotide sequences using saturation mutagenesis. Each of the three related parental nucleotide sequence was aligned in the computer to determine demarcation points, and 17 such points were identified. Once each demarcation point was determined, the system determined the sequence of the 18 different fragments that would make up each parental gene. Each fragment from the parental sequence had a unique 5' and 3' overhang so only genes in the proper order could be reassembled by the computer. Because there were 18 fragments and three parents, the system had a total of 18 X 3 = 54 total fragments to analyze. It is advantageous for the system to pre-ligate each of the fragments in a process in order to store datafiles corresponding to every possible combination of pre-ligated fragments. This allows the system to determine the proper quantities of each pre-ligated fragment at each step in the ligation reaction in order to generate a resulting progeny population that has a predetermined PDF. Thus, in this example, the computer determined and stored the following pre-ligated sequences into its memory for EACH parent sequence. Accordingly, the following pre-ligation method is carried out on each parent sequence, the resulting data is stored to the computer.

The nomenclature "F1_1" refers to the first fragment from the chosen parental sequence. The nomenclature "F1_5" corresponds, as shown below, to a dataset comprising a combination of the first, second, third, fourth and fifth fragments of the chosen parental sequence. Thus, the following listing illustrates that the system can generate a dataset that

stores every possible pre-ligated fragment for a given parent. This dataset is then used by the system to determine the proper quantities of each pre-ligated fragment to result in the desired final crossover population of progeny chimeric sequences.

5 **Listing of Pre-Ligation Dataset for a Parent Sequence having 18 fragments.**

F1_1 = F1_1
F1_2 = F1_1 + F2_2
F1_3 = F1_1 + F2_2 + F3_3
F1_4 = F1_1 + F2_2 + F3_3 + F4_4
10 F1_5 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5
F1_6 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6
F1_7 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7
F1_8 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8
F1_9 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9
15 F1_10 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10
F1_11 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11
F1_12 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12
F1_13 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13
F1_14 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 +
20 F14_14
F1_15 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 +
F14_14 + F15_15
F1_16 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 +
F14_14 + F15_15 + F16_16
25 F1_17 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 +
F14_14 + F15_15 + F16_16 + F17_17
F1_18 = F1_1 + F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 +
F14_14 + F15_15 + F16_16 + F17_17 + F18_18
F2_2 = F2_2
30 F2_3 = F2_2 + F3_3
F2_4 = F2_2 + F3_3 + F4_4
F2_5 = F2_2 + F3_3 + F4_4 + F5_5
F2_6 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6
F2_7 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7
35 F2_8 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8
F2_9 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9
F2_10 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10
F2_11 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11
F2_12 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12
40 F2_13 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13
F2_14 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 + F14_14
F2_15 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 + F14_14 +
F15_15
F2_16 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 + F14_14 +
45 F15_15 + F16_16
F2_17 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 + F14_14 +
F15_15 + F16_16 + F17_17
F2_18 = F2_2 + F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11 + F12_12 + F13_13 + F14_14 +
F15_15 + F16_16 + F17_17 + F18_18
50 F3_3 = F3_3
F3_4 = F3_3 + F4_4
F3_5 = F3_3 + F4_4 + F5_5
F3_6 = F3_3 + F4_4 + F5_5 + F6_6
F3_7 = F3_3 + F4_4 + F5_5 + F6_6 + F7_7
55 F3_8 = F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8
F3_9 = F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9
F3_10 = F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10
F3_11 = F3_3 + F4_4 + F5_5 + F6_6 + F7_7 + F8_8 + F9_9 + F10_10 + F11_11

$F3\_12 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12$

$F3\_13 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13$

$F3\_14 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14$

$F3\_15 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15$

5    $F3\_16 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16$

$F3\_17 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17$

$F3\_18 = F3\_3 + F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 +$

10    $F16\_16 + F17\_17 + F18\_18$

$F4\_4 = F4\_4$

$F4\_5 = F4\_4 + F5\_5$

$F4\_6 = F4\_4 + F5\_5 + F6\_6$

$F4\_7 = F4\_4 + F5\_5 + F6\_6 + F7\_7$

15    $F4\_8 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8$

$F4\_9 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9$

$F4\_10 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10$

$F4\_11 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11$

$F4\_12 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12$

20    $F4\_13 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13$

$F4\_14 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14$

$F4\_15 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15$

$F4\_16 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16$

25    $F4\_17 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17$

$F4\_18 = F4\_4 + F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17 + F18\_18$

$F5\_5 = F5\_5$

30    $F5\_6 = F5\_5 + F6\_6$

$F5\_7 = F5\_5 + F6\_6 + F7\_7$

$F5\_8 = F5\_5 + F6\_6 + F7\_7 + F8\_8$

$F5\_9 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9$

$F5\_10 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10$

35    $F5\_11 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11$

$F5\_12 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12$

$F5\_13 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13$

$F5\_14 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14$

$F5\_15 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15$

40    $F5\_16 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16$

$F5\_17 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17$

$F5\_18 = F5\_5 + F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17 + F18\_18$

45    $F6\_6 = F6\_6$

$F6\_7 = F6\_6 + F7\_7$

$F6\_8 = F6\_6 + F7\_7 + F8\_8$

$F6\_9 = F6\_6 + F7\_7 + F8\_8 + F9\_9$

$F6\_10 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10$

50    $F6\_11 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11$

$F6\_12 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12$

$F6\_13 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13$

$F6\_14 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14$

$F6\_15 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15$

55    $F6\_16 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16$

$F6\_17 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17$

$F6\_18 = F6\_6 + F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11 + F12\_12 + F13\_13 + F14\_14 + F15\_15 + F16\_16 + F17\_17 + F18\_18$

$F7\_7 = F7\_7$

60    $F7\_8 = F7\_7 + F8\_8$

$F7\_9 = F7\_7 + F8\_8 + F9\_9$

$F7\_10 = F7\_7 + F8\_8 + F9\_9 + F10\_10$

$F7\_11 = F7\_7 + F8\_8 + F9\_9 + F10\_10 + F11\_11$